Практикум на ЭВМ.
Задание №3.

Выполнили студенты
Факультета ВМК МГУ:
Семакова Екатерина
Пилипенко Александр
Доросева Екатерина

Анализ временного ряда и предсказание значения для последующих месяцев.

Содержание файла Readme.md:

---

1. Постановка задачи

Целью данного задания было изучение методов определения свойств временных рядов и ознакомление с различными моделями для предсказания значений. Также необходимо было провести оценку качества полученных моделей.

---

2. Алгоритм разрешения задачи

1. Проверка ряда на стационарность в широком смысле двумя способами:
    a. Проведение визуальной оценки, изображением ряда и скользящей статистики (среднее, стандартное отклонение).

   b. Проведение теста Дики - Фуллера.
   Далее оценка достоверности статистики.
2. Разложение временного ряда на тренд, сезональность, остаток в соответствии с аддитивной, мультипликативной моделями.
3. Проверка является ли временной ряд интегрированным порядка k. Если является, применение к нему модели ARIMA, подбором необходимых параметров с помощью функции автокорреляции и функции частичной автокорреляции. Отбор нескольких моделей. Предсказание значения для тестовой выборки. Отбор наилучшей модели с помощью информационного критерия Акаике.
4. Анализ получившихся результатов.

---

3. Теоретическое описание решения

В данной работе использовались следующие термины.

Временной ряд - совокупность наблюдений определенной величины (например, экономической) в различные моменты времени. Они задаются на фиксированном временном промежутке. Начало временного промежутка примем за 0, конец за T. Мы будем обозначать тестируемый временной ряд символом Y.

Временной ряд называется *строго стационарным (стационарным в узком смысле)*, если сдвиг во времени не меняет ни одной из функций плотности распределения. Следствием из определения будут независимость матожидания (обозначение E()) и дисперсии (обозначение D()) от времени.

Математическое ожидание — среднее значение случайной величины (распределение вероятностей стационарной случайной величины) при стремлении количества выборок или количества измерений её к бесконечности.

Дисперсия случайной величины — мера разброса значений случайной величины относительно её математического ожидания.

Имеет смысл рассмотреть ковариацию значений ряда в различные моменты времени - Cov[X(t1),X(t2)]. Она не зависит от сдвига времени вперед, а зависит только от разности моментов времени t_2-t_1 = T'.

Ковариация (корреляционный момент, ковариационный момент) — в теории вероятностей и математической статистике мера линейной зависимости двух случайных величин.

Автоковариационная функция - совокупность значений ковариаций при всевозможных значениях T' (Обозначение y). Является четной функцией для строго стационарного ряда.

Коэффициент корреляции - $Corr(T') = \frac{y(T')}{y(0)}$

График Corr(T') коэффициента корреляции носит название *коррелограммы*.

Слабая стационарность - стационарность с ослабленными условиями: $E[Y] = Const$, $D[Y] = Const$, $Corr(T') = Const$. Отличие от сильной стационарности заключается в том, что необязательно наличие четности коэффициента корреляции.

Трендом временного ряда называется изменение, определяющее общее направление развития ряда - рост, падение, неизменность.

Визуальная оценка в работе проводится следующим образом. Обратим внимание на наличие у графика тренда. Тренд на изменения есть - ряд нестационарный, тренда нет - ряд стационарен. Стоит отметить, что эта оценка довольно грубая, и во многих случаях из внешнего вида графика нельзя сказать, стационарен ряд или нет.

Авторегрессионная модель (AR) - модель временного ряда, в которой значение ряда y(t) линейно выражается через предыдущие значения этого же ряда. Если зависимость происходит только от последних p значений, то говорят, что задан авторегрессионный временной ряд порядка p (или AR(p)).

Процесс Скользящего Среднего (Moving Average - MA) - взвешенное среднее значений исходной функции. (Мы будем использовать простейшую форму взвешенного среднего - среднее арифметическое)

Также следует упомянуть области применения Скользящего Среднего:

- Сглаживание краткосрочных колебаний
- Выделение основных тенденций

Стандартное Отклонение - показатель среднего отклонения ряда от Скользящего Среднего (от среднего арифметического, в нашем случае).

Скользящее среднее вместе со Стандартным Отклонением составляют *Скользящие Статистики*.

В процессе решения данной задачи применяется *тест Дики-Фуллера*. Это метод анализа временных рядов на стационарность. Суть метода заключается в проверке ряда на наличие так называемых единичных корней.

Временной ряд имеет *единичный корень* (хотя бы один), если его первые разности образуют стационарный ряд.
(Обозначение *Y(t) ~ I(1), т.е. ΔY(t) = Y(t) - Y(t-1) ~ I(0),*
где Δ - разностный оператор, I(j) - означает, что ряд является интегрированным порядка j, I(0) - ряд стационарен)

Интегрированный временной ряд - нестационарный временной ряд, разности некоторого порядка от которого являются стационарным рядом. Временной ряд называется *интегрированным порядка k*, если разности ряда k-го порядка `Δ^k x(t)` являются стационарными, а разности меньшего порядка (и сам временной ряд соответственно) не являются стационарными рядами.

- *Δ Y(k) = Y(k+1) - Y(k)*
- *Δ^2 Y(k) = Δ Y(k+1) - Δ Y(k) = Y(k+2) - 2*Y(k+1) + Y(k)*
- *Δ^m Y(k) = Δ^m-1 Y(k+1) - Δ^m-1 Y(k), где Δ* - разностный оператор

Обозначают *Y(k) ~ I(k)* - интегрированный временной ряд порядка k

Фактически, тест Дики-Фуллера проверяет значение коэффициента `a` в авторегрессионном уравнении 1-го порядка - AR(1). Оно имеет вид:
*Y(t) = a * Y(t-1) + ε(t)*, где ε(t) - ошибка значения. В результате работы метода возможны 3 исхода:

- a=1 => есть единичные корни => стационарности нет
- |a|<1 => нет единичных корней => есть стационарность
- |a|>1 => не свойственно для временных рядов, которые встречаются в реальной жизни - требуется более сложный анализ.

Для удобства проведем преобразование уравнения:
*Y(t) = a * Y(t-1) + ε(t) =>*
*Y(t) - Y(t-1) = a * Y(t-1) - Y(t-1) + ε(t)*
*=> ΔY(t) = (a-1) * Y(t-1) + ε(t) =>*
*ΔY(t) = b * Y(t-1) + ε(t), где b = a-1*

- Основная гипотеза: H_0: b = 0 - процесс не стационарен
- Альтернативная гипотеза: H_1: b<0 - процесс стационарен

Проверка на наличие корней происходит следующим образом. В результате работы метода будут полученные значения:
q-value - достоверность статистики;
critical values - критические значения с уровнем значимости 5%.
(Данные параметры принимают значения от 0 до 1)

Достоверность статистики - мера уверенности в "истинности" результата. Чем она меньше, тем больше доверия. Как только достоверность статистики становится больше определенного параметра - `critical values`, то мы понимаем, что доверять нельзя и говорим, что единичные корни есть.

Уровень значимости - показывает (обычно в процентном выражении) степень отклонения от гипотезы. Грубо говоря, если наша достоверность превысила, скажем, 5% уровень значимости, то гипотеза отвергается, что говорит о том, что процесс не стационарен.

Если достоверность статистики больше критических значений, тогда единичный корень есть и ряд не стационарен. Иначе ряд является стационарным.

Также в работе использовались понятия сезонность и остаток.

Сезонность - периодические колебания уровней временного ряда внутри года

Остаток – величина, показывающая нерегулярную (не описываемую трендом и сезонностью) составляющую исходного ряда в определенном временном интервале. Фактически, остатком называется разница между предсказанным и наблюдаемым значением.

---

4. Описание используемых моделей

Модель, в которой временной ряд представлен как сумма перечисленных компонент, называется *аддитивной моделью* временного ряда. Общий вид аддитивной модели: $Y = T + S + E$, где T - тренд, S - сезональность, E - остаток.

Модель, в которой временной ряд представлен как произведение перечисленных компонент, называется *мультипликативной моделью* временного ряда. Общий вид мультипликативной модели: $Y = T * S * E$;

Аддитивную сезонность имеет смысл использовать, если амплитуда колебаний сезонности не меняется. Если амплитуда колебаний сезонности меняется (т.е. размах уменьшается или увеличивается), то более оптимально использовать мультипликативную сезонность**.**

4.1 Аддитивная модель

Для поиска сезональности аддитивной модели используется процесс скользящего среднего два раза последовательно, в результате получается центрированное скользящее среднее. Центрированное скользящее среднее (являющееся временным рядом) и будет нашей оценкой сезональности.

Поиск тренда аддитивной модели осуществляется с помощью метода наименьших квадратов и получается приближенный временной ряд, позволяющий найти тренд временного ряда.

Поиск остатка аддитивной модели получается из общего вида:
$E = Y - T - S$

4.2 Мультипликативная модель

Аналогично аддитивной модели найдем центрированное скользящее

среднее. Сезональность (S) = Временной ряд (Y) / Центрированное скользящее среднее. Также аналогично аддитивной модели находим тренд и остаток из общего вида:
*E = Y / (S * T)*

4.3 Модель ARIMA

ARIMA (autoregressive integrated moving average model) - интегрированная модель авторегрессии скользящего среднего - модель анализа временных рядов. Это расширение моделей ARMA для нестационарных временных рядов.

ARMA (AutoRegressive Moving Average model) - математическая модель, используемая для анализа и прогнозирования стационарных временных рядов. Объединяет 2 более простые модели: авторегрессии (AR) и скользящего среднего (MA).

AR (autoregressive model) - авторегрессионная модель временных рядов, в которой значение временного ряда в данный момент зависит от предыдущих значений этого же ряда.

MA (moving average model) - модель скользящего среднего, в которой моделируемый уровень временного ряда можно представить как линейную функцию прошлых ошибок, т.е. разностей между прошлыми фактическими и теоретическими уровнями.

Алгоритм построения модели ARMA заключается в поиске коэффициентов p, q - порядков для моделей AR(p) и MA(q). Это позволит построить функцию автокорреляции и функцию частичной автокорреляции.

Таким образом, построение ARIMA зависит от 3 параметров: ARIMA(p, d, q), где p - порядок AR(p), d - порядок интегрированности, q - порядок MA(q).
ARIMA(p, d, q) = AR(p) + MA(q) ~ I(d)

---

5. Описание подсчета коэффициента детерминации R_2

Коэффициент детерминации (R_2) - доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью зависимости. Более точно — это единица минус доля необъяснённой дисперсии (дисперсии случайной ошибки модели, или условной по факторам дисперсии зависимой переменной) в дисперсии зависимой переменной.

Лучший результат, который может дать R_2 score — это 1.0. Однако возможны как отрицательные значения, так и значения, превышающие 1.0. Это говорит об очень большой неточности предсказания модели.

Информационный критерий - применяемая в эконометрике мера относительного качества эконометрических моделей, учитывающая степень "подгонки" модели под данные с корректировкой на используемое количество оцениваемых параметров. Т.е.

критерии основаны на некотором компромиссе между точностью и сложностью модели. Критерии различаются тем, как они обеспечивают этот баланс. информационные модели используются исключительно для сравнения моделей между собой, без содержательной интерпретации значений этих критериев. Обычно чем меньше значения критериев, тем выше относительное качество модели.

AIC (an information criterion) - информационный критерий Акаике – критерий, применяющийся исключительно для выбора из нескольких статистических моделей.

AIC = 2k - 2ln(L), где где k — число параметров в статистической модели, и L — максимизированное значение функции правдоподобия модели.

---

6. Описание решения задачи

1. Чтение данных из `training.csv`
   - Чтение производится `командой read_csv('training.csv')`
     - Библиотека `pandas`

2. Проверка на стационарность
   - Визуальная оценка
     - `ts.rolling().mean()` - Скользящее среднее
       - Возвращает скользящее среднее
       - Библиотека `pandas`
     - `ts.rolling().std()` - Стандартное отклонение
       - Возвращает стандартное отклонение
       - Библиотека `pandas`
     - Строим вывод о стационарности ряда по визуальной оценке
   - Тест Дики-Фуллера
     - `sm.tsa.adfuller(ts)` - проводит тест Дики-Фуллера
     - ts - временной ряд
     - Возвращает массив с данными
       - 0 => adf
       - 4 => critical values
     - Библиотека `statsmodels`
     - Анализируем полученные из функции параметры
     - Строим вывод о стационарности данного ряда

3. Разложение временного ряда на тренд, сезональность остаток в соответствии с аддитивной и мультипликативной моделями
   - `seasonal_decompose(data.Value, model)`
     - data.Value - столбец значений временного ряда
     - model - 'additive' или 'multiplicate'
     - Возвращает decompose
       - * decompose.trend - тренд исходного ряда
       - * decompose.resid - остаток исходного ряда

        - * decompose.seasonal - сезонность исходного ряда
    - Библиотека `statsmodels`
- Строим вывод на основании полученных результатов

4. Поиск коэффициента интегрируемости ряда
- `order - коэффициент интегрированности ряда

5. Подбираем нужные параметры с помощью функции автокорреляции и функции частичной автокорреляции
- `plot_acf(y, lags, ax)` - находит и отрисовывает автокорреляцию временного ряда
    - y - временной ряд
    - lags - число лагов для автокорреляции
        - ax - параметр для построения подграфика
    - Зачем нужна? Помогает найти порядок q модели MA(q) для построения модели ARIMA(p, d, q)
    - Возвращает массив с значениями функции
    - Библиотека `statsmodels`
- `pacf(y, lags, ax)` - находит частичную автокорреляцию временного ряда
    - y - временной ряд
    - lags - число лагов для частичной автокорреляции
        - ax - параметр для построения подграфика
    - Зачем нужна? Помогает найти порядок p модели AR(p) для построения модели ARIMA(p, d, q)
    - Возвращает массив с значениями функции
    - Библиотека `statsmodels`
- Функции автокорреляции и частичной автокорреляции - дискретные, число различных значений равно лагу (в обоих случаях по 20). Соответствующие коэффициенты определяются по правилу - берется целая часть наибольшего значения. В обоих случаях наибольшее значение равно 1.0 и достигается оно на нулевом элементе (в обоих случаях).

6. ARIMA-модель
- `ARIMA(data, order=(p, d, q))` - строит модель ARIMA на основе временного ряда ts
    - data - временной ряд
    - order - порядки для модели
    - Возвращает объект ARIMA
    - Библиотека `statsmodels`
- `model.fit()` - Прогоняем ARIMA-модель через фильтр Калмана
    - Фильтр Калмана - мощнейший инструмент фильтрации данных. При фильтрации используется информация о физике самого явления.
    - Возвращает объект ARIMAResults - содержит также и результаты предсказаний, которые можно выводить и использовать в дальнейших прогнозах
    - Метод класса ARIMA
- `model.predict(start='1989-01-01', end='1993-12-01', typ='levels', dynamic=True)`

- строим предсказание на тестовой выборке
    - start - определяет с какой позиции начать предсказывать
    - end - определяет до какой позиции предсказывать
    - typ='linear' - строит предсказание по уровням временного ряда
    - dynamic=True - повышает точность предсказания
    - Возвращает предсказанные значения
    - Библиотека `statsmodels`

7. R_2 score
- `$r_2$_score(y_true, y_pred)` - подсчитывает r_2 score между реальной выборкой и предсказанной
    - y_true - реальные значения
    - y_pred - предсказанные значения
    - Библиотека `sklearn.metrics`
    - Возвращает r_2 score

8. AIC
- `model.aic` - выводит значение критерия - вычисляется в процессе построения статистической модели

---

7. Инструкции по запуску

Cell -> Run All

7.1 Необходимое ПО

Anaconda navigator

7.2 Библиотеки

1) Pandas - для работы с временными рядами
2) Numpy - для работы с массивами
3) Matplotlib.pylab - для построения графиков функций
4) Statsmodels
    a) Используются статистические функции
    b) Для построения модели ARIMA
5) Sklearn.metrics - Для метрики R_2 score

7.3 Программы

Jupyter Notebook